Deliverable

D1.12 4 workshops, videoconferences and jamborees for hands-on discussion on diagnostic hypothesis	
Version | Status	V2 | final
Work package	WP1
Lead beneficiary	INSERM-Orphanet (Ana Rath)
Due date	31.12.2020 (M36)
Date of preparation	20.06.2023
Target Dissemination Level	Public
Author(s)	P9a INSERM-Orphanet: Ana Rath, David Lagorce, Oscar Hognat, Emeline Lebreton, Annie Olry, Marc Hanauer P7 CNAG-CRG: Leslie Matalonga
Reviewed by	Rebecca Schüle (EKUT), Ana Topf (UNEW)
Approved by	Holm Graessner (EKUT)

The Solve-RD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 779257.

Explanation according to GA Annex I:

Workshops/jamborees and videoconferences for ultra-rare and « unsolvable symptoms » will be organized in order to engage ERNs in the phenotypic delineation of these rare disease (RD), and to potentially elaborate and discuss diagnostic hypothesis derived from the ontological approach for unsolvable cases.

Abstract

A series of 4 jamborees was organised in order to discuss the results of a proposed methodology based on phenotypic similarity calculations and reanalysis of genomic information provided by ERNs' clinicians on GPAP. Jamborees were organised by selecting a solved case presented by the data submitter and further discussed with regard to the consistency with a known rare disease (RD), with the ultimate goal to give a clinical diagnosis or, if inconsistent, to consider the emergence of a new RD by clustering with similar cases. In addition, a cascade of similarity calculations allowed for reanalysing unsolved cases for candidate genes and discussing these cases with the clinicians. Even if an estimation of the overall performance of this approach cannot be reported, it allowed to correctly identify variants in solved cases, and to raise new hypotheses deserving further investigation in at least 7 unsolved cases, and other cases are still to be discussed with the clinicians as there was not enough time to present them all. Out of these promising results, the methodology was validated and next steps generalising and standardising the method were decided. Selected unsolved and solved but yet clinically undiagnosed cases will be the object of the next series of jamborees.

Introduction

Currently, up to 50% of RD patients remain undiagnosed even after having undergone in depth molecular and clinical analysis. Similarly, for up to 30% of known, clinically defined rare diseases entities, the genetic background remains elusive. Some of these represent new genes. In other instances such unsolved RD are due to variation in known disease genes but with a novel function, mutation type or mechanism.

One of the main recognized challenges identifying the rare diagnosis for a patient is the lack of collection and exploitation of good-quality, standardised phenomics data[1][2][3].

WP1 aims at collecting phenomic and genomic data from unsolved RD cases, to share this information in a structured, standardized way, and to pool unsolved cases in a computable, ontological format together with all known RD in order to raise diagnostic hypotheses to be submitted for further investigation in the project. These hypotheses could lead to the identification of a new, formerly undescribed disease, or to determine that the patient suffers from a known disease with likely new or unreported manifestations. In both situations, the final goal is to return a diagnosis to the patient, by giving a name to his/her disease, and allow for visibility in Health Information Systems by attributing an ORPHAcode[4].

[1] Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. Nat Rev Genet. 2010 Dec;11(12):855-66. doi: 10.1038/nrg2897. PMID: 21085204.

[2] Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, Brookes AJ, Brudno M, Carracedo A, den Dunnen JT, Dyke SOM, Estivill X, Goldblatt J, Gonthier C, Groft SC, Gut I, Hamosh A, Hieter P, Höhn S, Hurles ME, Kaufmann P, Knoppers BM, Krischer JP, Macek M Jr, Matthijs G, Olry A, Parker S, Paschall J, Philippakis AA, Rehm HL, Robinson PN, Sham PC, Stefanov R, Taruscio D, Unni D, Vanstone MR, Zhang F, Brunner H, Bamshad MJ, Lochmüller H. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. Am J Hum Genet. 2017 May 4;100(5):695-705. doi: 10.1016/j.ajhg.2017.04.003. PMID: 28475856

[3] Deans AR, Lewis SE, Huala E, Anzaldo SS, Ashburner M, Balhoff JP, Blackburn DC, Blake JA, Burleigh JG, Chanet B, Cooper LD, Courtot M, Csösz S, Cui H, Dahdul W, Das S, Dececchi TA, Dettai A, Diogo R, Druzinsky RE, Dumontier M, Franz NM, Friedrich F, Gkoutos GV, Haendel M, Harmon LJ, Hayamizu TF, He Y, Hines HM, Ibrahim N, Jackson LM, Jaiswal P, James-Zorn C, Köhler S, Lecointre G, Lapp H, Lawrence CJ, Le Novère N, Lundberg JG, Macklin J, Mast AR, Midford PE, Mikó I, Mungall CJ, Oellrich A, Osumi-Sutherland D, Parkinson H, Ramírez MJ, Richter S, Robinson PN, Ruttenberg A, Schulz KS, Segerdell E, Seltmann KC, Sharkey MJ, Smith AD, Smith B, Specht CD, Squires RB, Thacker RW, Thessen A, Fernandez-Triana J, Vihinen M, Vize PD, Vogt L, Wall CE, Walls RL, Westerfeld M, Wharton RA, Wirkner CS, Woolley JB, Yoder MJ, Zorn AM, Mabee P. Finding our way through phenotypes. PLoS Biol. 2015 Jan 6;13(1):e1002033. doi: 10.1371/journal.pbio.1002033. PMID: 25562316; PMCID: PMC4285398.

[4] RD-CODE. Existing experiences and guidelines about coding of undiagnosed RD patients *http://www.rd-code.eu/existing-experiences-and-guidelines-about-coding-of-undiagnosed-rd-patients*

With the objective to delineate phenotypically known and new RD using standard disease and phenotype ontologies, phenotype-sharing expert-sessions, further called "phenotype jamborees" have been planned.

As a preparatory phase for these phenotype jamborees, a workshop has been organised in July 10th 2020, gathering over 40 people, with the aims of increasing ERNs' partners' knowledge about WP1 workflow, to increase awareness on the importance of good-quality deep phenotyping, and finally, to collect clinicians' feedback on how to best provide similarity results back to them.

Based on the lessons-learned from this preliminary workflow, a methodology has been designed and a first series of phenotypic jamborees was organised based on solved cases as a starting point for diagnostic hypothesis for unsolved cases.

Report

1. Phenotype jamboree's objectives

The first series of jamborees were called **"Giving a disease name to solved cases and trying to solve the unsolved"**. The main objectives of the jamborees were:

- to understand if the genetically **solved cases** define new disorders or better delineates known ones, so giving a name to patient's disease
- to validate the methodology based on phenotypic similarity to raise diagnostic hypotheses for **unsolved cases**
- to allow for **establishing standardized protocols** based on the methodology
- to define the cases to illustrate the methodology in a publication **if results are consistent.**

2. Methodology

In order to achieve jamborees' objectives, a methodology was set up that takes a case tagged as "solved" in the GPAP platform and for which both the phenotypes and genotype are annotated, as the triggering case; then to run a cascade of phenotypic similarity calculations and to reanalyse the genomic information present in GPAP for both solved and unsolved cases based on phenotypic similarity results.

Phenotypic similarity calculations were carried out using the Resnik Symmetric algorithm[5,6,7,8] and the 50 first ranked results were retrieved, irrespective of the similarity score between them. This differs compared to the methodology previously reported in deliverable (see results section below).

Genomics data reanalysis was performed by filtering data in GPAP database for the genes selected based on the similarity algorithms (see below), and looking for variants for which

[5] Bauer S, Köhler S, Schulz MH, Robinson PN. Bayesian ontology querying for accurate and noise-tolerant semantic searches. Bioinformatics. 2012 Oct 1;28(19):2502-8. doi: 10.1093/bioinformatics/bts471.

[6] Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, Mundlos C, Horn D, Mundlos S, Robinson PN. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009 Oct;85(4):457-64. doi: 10.1016/j.ajhg.2009.09.003.

[7] Using information content to evaluate semantic similarity in a taxonomy. Resnik, P. (1995) Proc 14th Int Joint Conf Artificial Intelligence.

[8] Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Comput Biol. 2009 Jul;5(7):e1000443. doi: 10.1371/journal.pcbi.1000443.

population frequency is reported to be 0.01 in gnomAD and internal frequency and is annotated as having a high (truncating) or moderate (amino acid change) at the protein level according to **SNPeff.** The overall reanalysis approach was performed in a programmatic way by using the RD-Connect GPAP API (Matalonga et al., manuscript under review at EJHG). Resulting variants were submitted to WP1 and DITF members for final evaluation and discussion during the jamboree.

The following figures illustrate the **3-step overall workflow** used to prepare the jamborees:

Step A: Working out the solved cases (see Figure 1).

A1: The clinical consistency between the solved case phenotype and the corresponding RD (in Orphanet Rare Disease Ontology [ORDO]) caused by mutations in the same gene as the solved case is discussed.

A2: In case of clinical inconsistency between the case and the known RD, the similarity algorithm is run in order to detect the first 50 similar disorders (using ORDO). Their causative genes are then re-analysed based on data present in GPAP to eventually detect candidate variants.

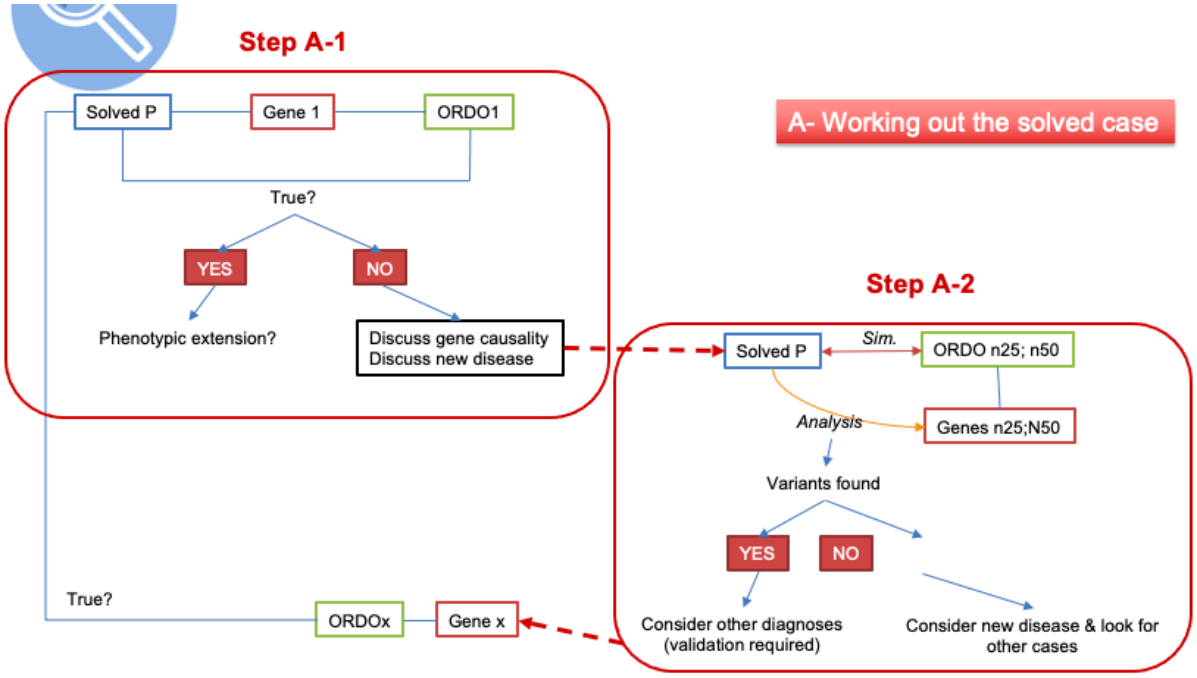

Figure 1: Step A – Working out the solved cases.

Step B: Finding the unsolved cases similar to the solved case (see Figure 2).

B1. Genomics information of the top 50 cases phenotypically similar to the solved case is reanalysed for the gene causative of the solved case.

B2. For cases in which no variant in the candidate gene is found, then each unsolved case is reanalysed for the genes causative of its top 50 most similar diseases (using ORDO).

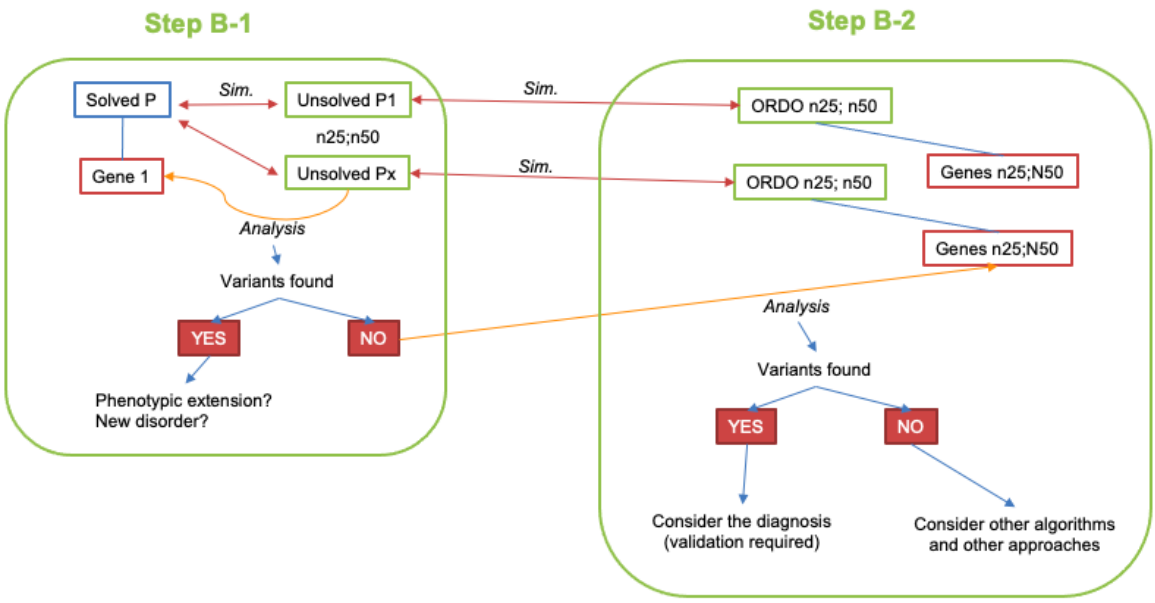

Figure 2: Step B – Finding the unsolved cases similar to the solved case.

Step C: Working out the unsolved cases similar to the RD related to the solved case (see Figure 3).

C1. Genomics information of the top 50 cases phenotypically similar to the RD (in ORDO) caused by the gene involved in the solved case is reanalysed looking for variants in this gene.

C2. For cases in which no variant in the candidate gene is found, then each unsolved case is reanalysed for the genes causative of its top 50 most similar diseases (using ORDO).

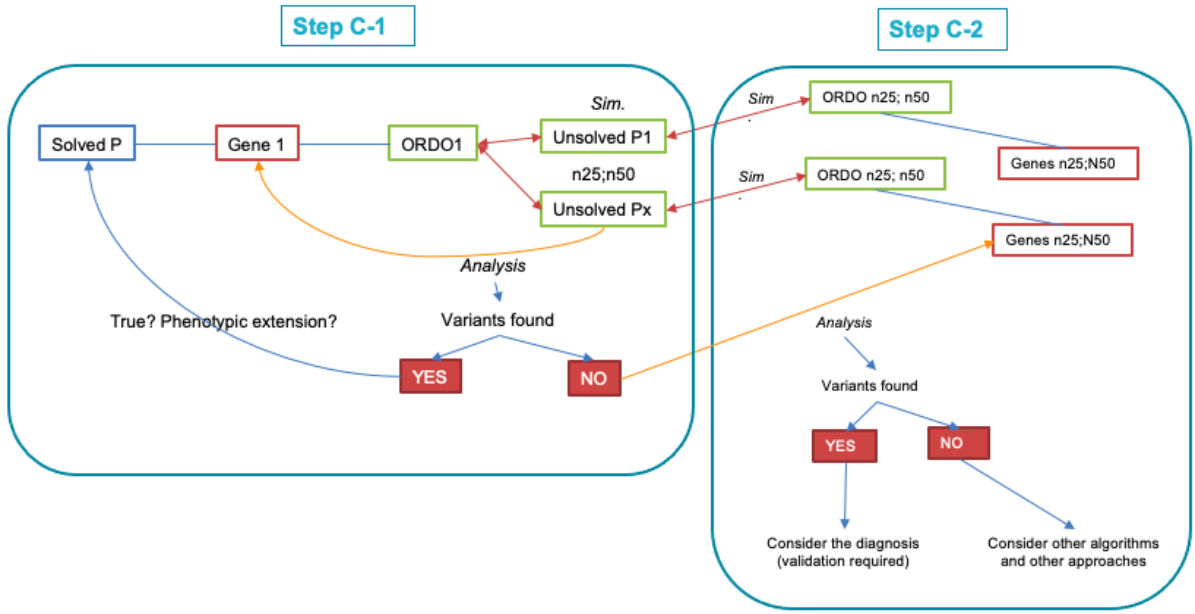

Figure 3: Step C – Working out the unsolved cases similar to the RD related to the solved case.

Global similarity results were presented as Cytoscape[9] networks produced out of Rare Disease Case Ontology (RDCO) and highlighted cases were further discussed based on related Human Phenotype Ontology (HPO) terms comparison and genomics reanalysis results.

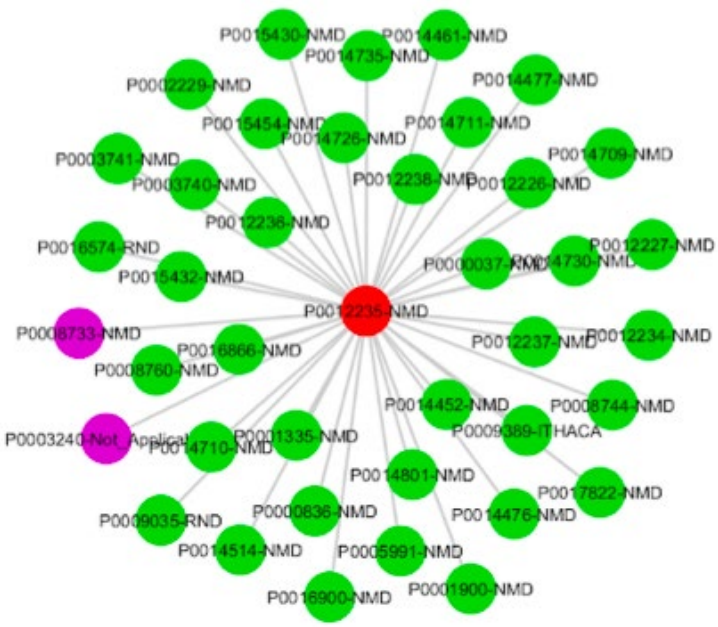

Figure 4: Example of Cytoscape network representation on unsolved similar cases (green dots) to a solved case (red dot). Cases with candidate variants are highlighted (violet dots).

Case selection: In order to perform the 4 planned jamborees, 4 solved cases were selected to explore different situations:

- A case of a mostly pure hereditary spastic paraplegia (but potentially presenting with a more variable phenotype) with peculiar characteristics according to the HPO annotations in GPAP (SPAST-related phenotype)
- A case which HPO annotations in GPAP did not allow to match with the gene-related ORDO disease when running similarity calculations (TBL1XR1-related phenotype)
- A case which causative gene is related to >1 RD in ORDO (CASQ1-related phenotype)
- A case which phenotypic description nicely matched the related ORDO RD (KIF5A-related phenotype).

3. Results

Similarity algorithm performance.

Resnik symmetric has been formally selected in the project after performance comparison of eight different algorithms (see Figure 5 below) based on the first 59 solved cases listed at the very beginning of the project (also see deliverable report D1.10).

[9] Cytoscape is an open source software platform for visualizing complex networks and integrating these with any type of attribute data. https://cytoscape.org/

Algorithm	Uses Frequency Information	Based on	Reference(s)
Resnik (symmetric)	No	Information content (IC)	Resnik et al. (1995), Pesquita et al. (2009), Köhler et al. (2009)
Resnik (asymmetric)	No		
Bernoulli with Grid (BOQA-like)	No	Bayesian approach & Inherent IC	Bauer et al. (2012)
PhenoDigm	No	Information content (IC)	Smedley et al. (2013)
Jaccard	No	HPO subclass hierarchy	Pesquita et al. (2009), Köhler (2018)
Jaccard (weighted)	Yes		
Cosine	No	HPO subclass hierarchy	Pesquita et al. (2009), Köhler (2018)
Cosine (weighted)	Yes		

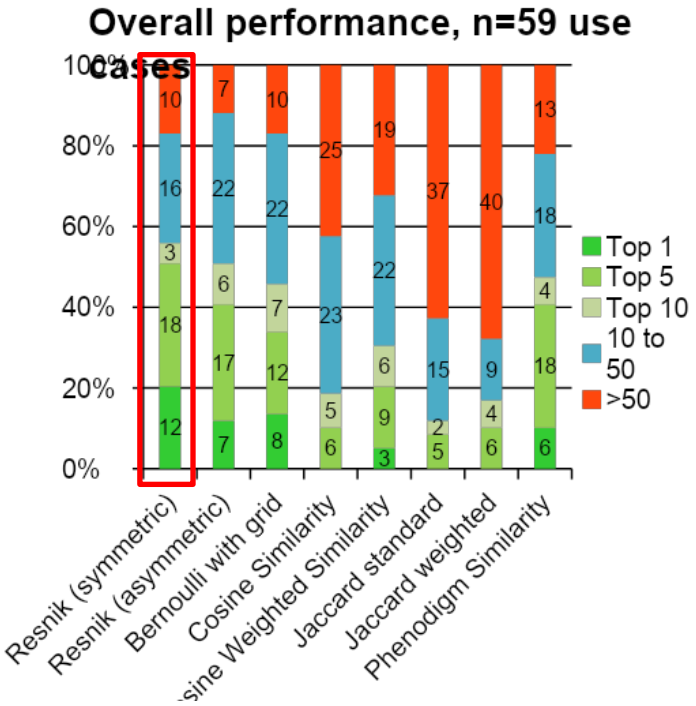

Figure 5: Performance comparison of eight different algorithms based on the first 59 solved cases.

Similarity computations rely on phenotypic annotations, by using Human Phenotype Ontology, between cases themselves (where 50% of the 5041 annotated cases contain 1 to 12 annotations (median = 8) top left in Figure 6) and/or cases and ORPHA entities (recent analysis were performed on R, showing that 75% of the 3961 annotated ORPHA contain 2 to 33 annotations (median = 21) and where less than 5% of these ORPHA are beyond 64 annotations so they can be considered as outliers bottom left in Figure 6). Based on the performance statistics of 465 solved cases for which genes are reported in GPAP, it was observed that most similar results are found at scores as low as 0.45-0.55 with the right diagnosis being found in the first 30 best similar results, then stabilising in a plateau. Therefore, an optimal capture of the right diagnosis is achieved when considering the first 50 results. It can be explained by the huge discrepancy of number of HPO terms used between cases in GPAP and disease annotations in Orphanet.

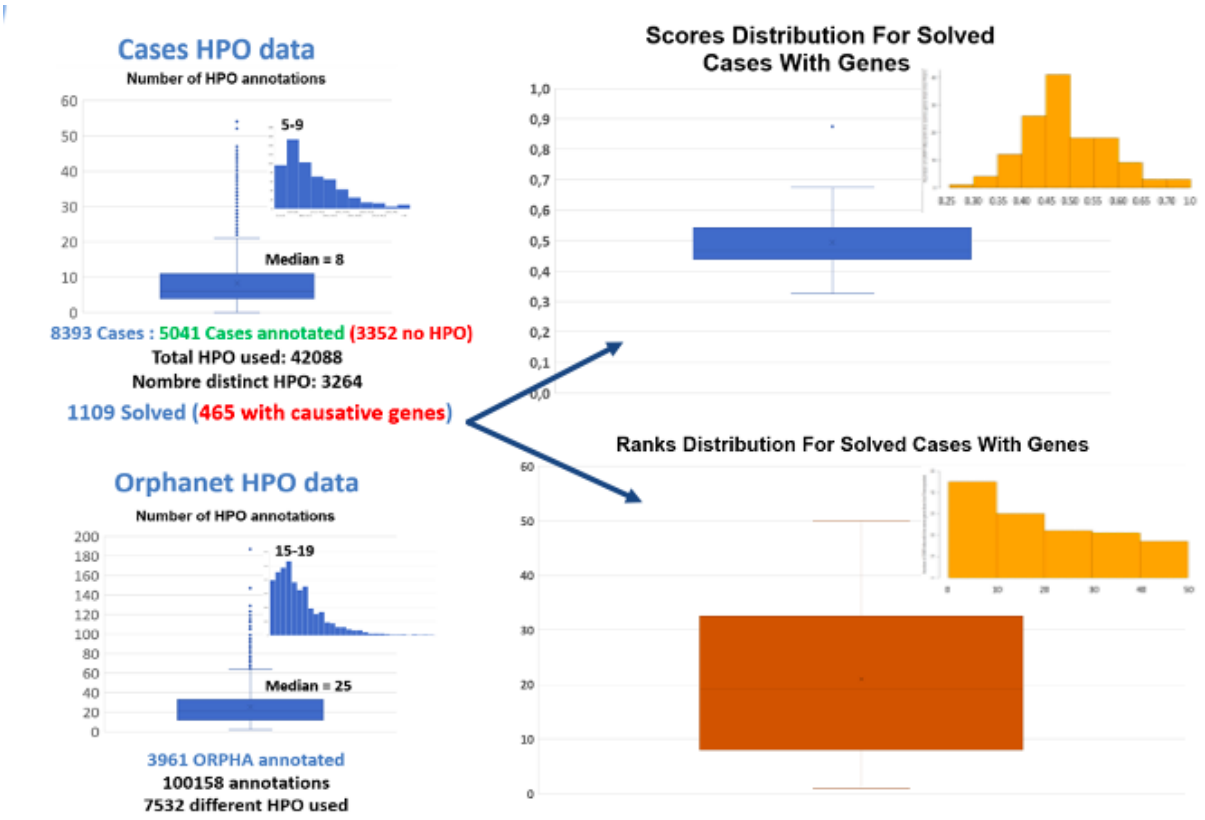

Figure 6: Similarity computations between cases themselves (top left) and/or cases and ORPHA entities (bottom left).

Jamborees' results

Solved cases submitters were invited to present and discuss their cases. A list of unsolved cases for which candidate variants have been found through the workflow steps were also invited to bring their cases to discussion. For those for who this was not possible, anonymised data was sent by mail and further exchanges are ongoing.

Jamborees took place on 18, 19, 20 and 25 January 2021. They were attended by around 17 people each (18, 16, 19, and 17, respectively). Jamborees were held online because of the COVID pandemics. In order to keep the case discussions private, recordings were stored locally to allow delivering this report, then destroyed.

The Agenda was as follows:

- Background information and presentation of the methodology by WP1 partners
- Solved case presentation by the clinician
- Discussion on Step A
- Unsolved cases work-up (steps B and C): presentation of the results by WP1 partners and discussion of selected cases by the clinicians.
- Overall conclusions and feedback

Summary of the results

Step A: Working out the solved cases (arrows indicate the actions to be taken)

Solved cases	A1	A2
SPAST	Related RD (ORPHA:100985) found at rank 28.	Look for homozygous CNVs: good candidate: a homozygous deletion in

	Some phenotypic traits might not be explained by the SPAST variant.	LDHA (11:18422383-18422557) - GSD due to lactate dehydrogenase M-subunit deficiency (ORPHA:284426) ➔ Further investigation by clinicians
TBL1XR1	Related RD NOT found by Resnik sym. Case HPO annotations not exhaustive but the case is likely ORPHA:487825 Pierpont syndrome ➔ Clinician to deep-phenotype and A1 step to be run again	Not informative
CASQ1	ORPHA:88635 found at rank 19. Discussion of the possible 2 related diagnoses led to the conclusion that other genes could be involved in the phenotype.	Step A2 to be performed in consequence ➔ Case discussion to be continued after running step A2
KIF5A	ORPHA:100991 found at rank 11 Case-variant-matching RD are consistent	Not informative

Steps B and C: Working out the unsolved cases (arrows indicate the actions to be taken)

Triggering solved cases	**B1**	**B2**	**C1**	**C2**
SPAST	SPAST variant candidate found in 1 unsolved case that could explain the case, which however remains intriguing because not segregating in symptomatic offspring. **Case remains unsolved**	9 variants found, of which 2 candidate variants. **1 interesting variant:** ➔ Deserves further investigation	**2 variants found** in SPAST explaining 2 solved cases (not reported as solved in GPAP): **Positive confirmation of the method** Variants in 3 other solved cases were correctly identified	16 candidate variants ➔ Ongoing mail exchanges with the clinicians (cases couldn't be presented)
TBL1XR1	Likely pathogenic variant in a typical ALS case: Clinically inconsistent	42 variants /972 genes studied, of which 4 candidates, but none explanatory of unsolved the cases	N/A (no unsolved cases similar to corresponding RD)	N/A
CASQ1	No *CASQ1* variants found	2 candidate variants found: ➔ 2 cases to be followed up with clinicians	No *CASQ1* variants found in unsolved cases in this step	Candidate variants found for 10 cases ➔ 4 cases to be followed up with clinicians
KIF5A	1 candidate variant found ➔ Possibly solving the case: to be further studied	1 candidate variant found on a case discussed in *SPAST* jamboree and	Likely pathogenic KIF5A variants found in 4 cases ➔ Variants to be reclassified as	Candidate variants identified for 6 cases of which: - 1 already solved: positive control for the procedure

		deserving further investigations	no in the *KIF5A* motor region	- 1 possibly solving the case ➔ to be further studied ; deep-phenotyping to be revised - 1 case deserving further investigation: ➔ needs phenotype reannotation and re-run the algorithms ➔ 4 other cases need follow-up with clinicians

Conclusions and further steps

Implementing a methodology based on phenotype similarity calculations as proposed here is promising for it allows to detect candidate variants and CNVs that could have been missed during the first investigations, or to help explaining phenotypes that are not fully consistent with a suspected diagnosis. The approach encounters, however, several limitations:

- Phenotypic annotations are frequently scarce: these jamborees helped in increasing the awareness on how good quality deep phenotyping is important for improving the results of this kind of approach. It is worth to note that AI-based approaches are also dependent on quality of data to feed automated reasoning algorithms: our semi-automated approach helped attendees understand the dependency of these approaches on the quality of clinical data. However, producing phenotype annotations is burdensome for clinicians and there is room for testing other approaches (i.e. Natural Language Processing (NLP) and entity recognition in clinical narratives)
- Phenomics and genomics data in the database is sometimes old and not taking into account the clinical evolution (or even the resolution) of cases.
- Variant classification in reference databases are not always accurate, underlying the effort needed in collaborative, expert reviewed, gene curation approaches.

Interestingly, 2 out of 4 solved cases did not completely match the known RD associated with the causative gene, failing accurately naming the disease, and thus needing further investigation (second gene?, regulatory sequences?).

As all the cases for which new data was generated could not be discussed at the jamborees and are currently being discussed remotely with clinicians, a final performance result cannot be derived. Furthermore, the number of cases selected to run these proof-of-principle jamborees is too small. Nevertheless, new interesting diagnostic hypotheses resulted from it, deserving further investigations, and positive control cases were found. It was then decided to standardise the process in order to:

- Expand the approach from all solved cases in GPAP;
- Standardise the approach for unsolved cases, based on B2 and C2 steps in the methodology presented here. Unsolved cases should then be reanalysed looking for variants in their similar solved cases and their similar RD in ORDO.

The ways to automatise (including shared and/or collaborative process between Orphanet and GPAP) the reanalysis of variants in GPAP will therefore be studied, as it was done manually for the purpose of the jamborees.

Following suggestions from the clinicians, results from a standardised approach should be submitted to them in order to select cases for future jamborees together with them, in order to prepare the data and, eventually, improve the quality of phenotypic annotations: a series of 1-hour distant phenotype jamborees could then be proposed based on unsolved cases. Cross-ERNs similar cases found by our approach are another criterion to select cases to bring to discussions: this will be also taken into account.

Furthermore, DITF is invited to propose jamborees on solved cases with unusual phenotypic abnormalities for which the methodology presented here could be applied.

The results described in this deliverable report have been submitted for publication to the European Journal of Human Genetics: Emeline Lebreton *et al*. “Phenotypic similarity-based approach for variant prioritization for unsolved rare disease: a preliminary methodological report”.

The manuscript is currently being reviewed. The preprint is available from Research Square via https://doi.org/10.21203/rs.3.rs-2948814/v1.